WEBSOM - Self-Organizing Maps of Document Collections

Timo Honkela, Samuel Kaski, Krista Lagus, and Teuvo Kohonen

Helsinki University of Technology
Neural Networks Research Centre
P.O.Box 2200, FIN-02015 HUT, Finland

Abstract

Searching for relevant text documents has traditionally been based on keywords and Boolean expressions of them. Often the search results show high recall and low precision, or vice versa. Considerable efforts have been made to develop alternative methods, but their practical applicability has been low. Powerful methods are needed for the exploration of miscellaneous document collections. The WEBSOM method organizes a document collection on a map display that provides an overview of the collection and facilitates interactive browsing. Interesting documents can be retrieved by a content addressable search of interesting map locations. The interesting locations could also be marked as filters for collecting interesting new documents.

1 Introduction

In the WEBSOM project, insightful methods have been developed for information retrieval based on the Self-Organizing Map (SOM) [3]. The WEBSOM is an explorative full-text information retrieval method and a browsing tool [1, 2, 5]. In the WEBSOM similar documents become mapped close to each other on the map, like the books on the shelves of a well-organized library. The self-organized document map offers a general idea of the underlying document space. The user may view any area of the map in detail by simply clicking the map image with the mouse. The WEBSOM browsing interface (cf. Fig. 1) is implemented as a set of HTML documents that can be viewed using a graphical WWW browser.

The potential of the WEBSOM method has been demonstrated in case studies where articles from Usenet newsgroups have been organized. Some demonstrations of the WEBSOM are also available at the WWW address `http://websom.hut.fi/websom/`.

2 On the WEBSOM Method

The problem addressed by the WEBSOM method is to automatically order, or organize, arbitrary free-form textual document collections to enable their easier browsing and exploration.

Before ordering the documents they must be encoded; this is a crucial step since the ordering depends on the chosen encoding scheme. In principle, a document might be encoded as a histogram of its words, whereby for computational reasons the order of the words is neglected. The computational burden would still, however, be orders of magnitude too large with the vast vocabularies used for automatic full-text analysis. An additional problem with the word histograms is that two words with similar meaning contribute to the histogram as differently as any other pair of words. In a useful full-text analysis method synonymous expressions should, however, be encoded similarly.

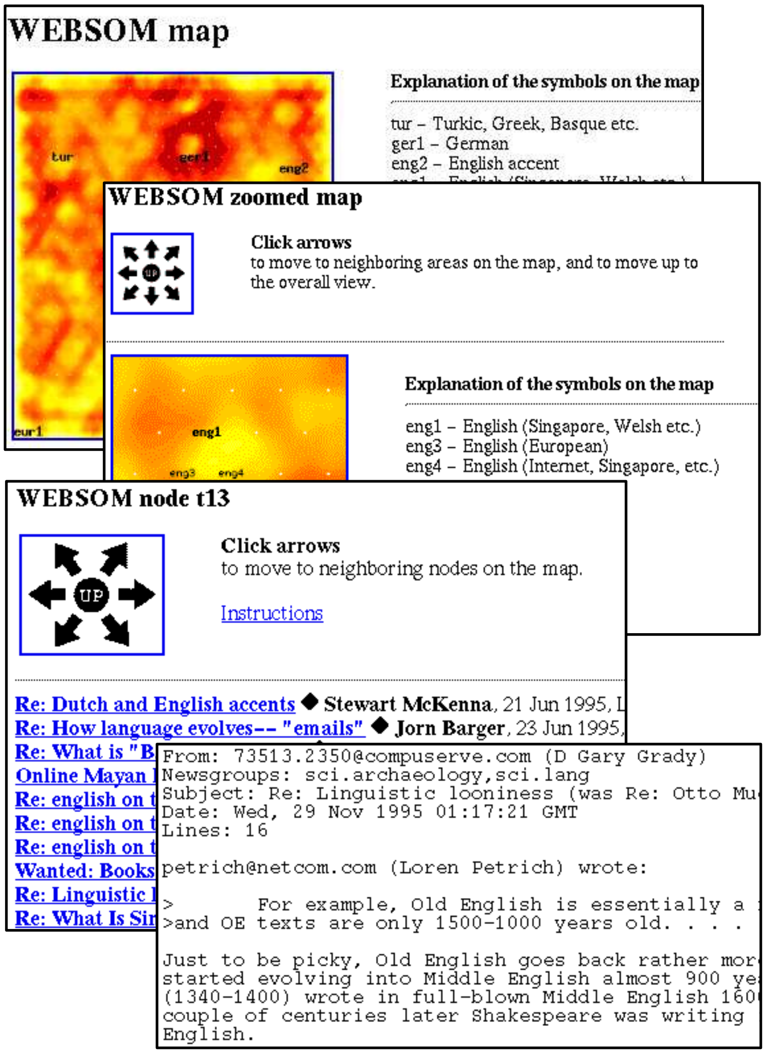

Figure 1: The different view levels in the WEBSOM browsing interface: whole map, zoomed map, map node, and single document, in the order of increasing detail. Moving between the levels or to neighboring areas on the same level is done by mouse clicks on the images or on the document links. Once an interesting area on the map has been found, one may use the arrow images to proceed to explore neighboring areas containing related documents. This can be contrasted with the traditional information retrieval techniques, where the users cannot know whether there are considerable numbers of relevant documents just "outside" the search results.

Since it is not currently feasible to incorporate references to real-life experience of word meanings in a text analysis method, the remaining alternative is to use the statistics of the textual contexts of words to provide information on their relatedness. It has turned out that the size of the word histograms can be reduced to a fraction with the so-called "self-organizing semantic maps" [8]. At the same time the semantic similarity of the words can be taken into account in encoding the documents. The basic processing architecture of the WEBSOM method is presented in Fig. 2.

The following sections present the details of actual processing in WEBSOM method as it has typically been applied in the case studies, in processing of the Usenet newsgroup articles. The main phases include preprocessing of the input, formation of the word category map, and, finally,

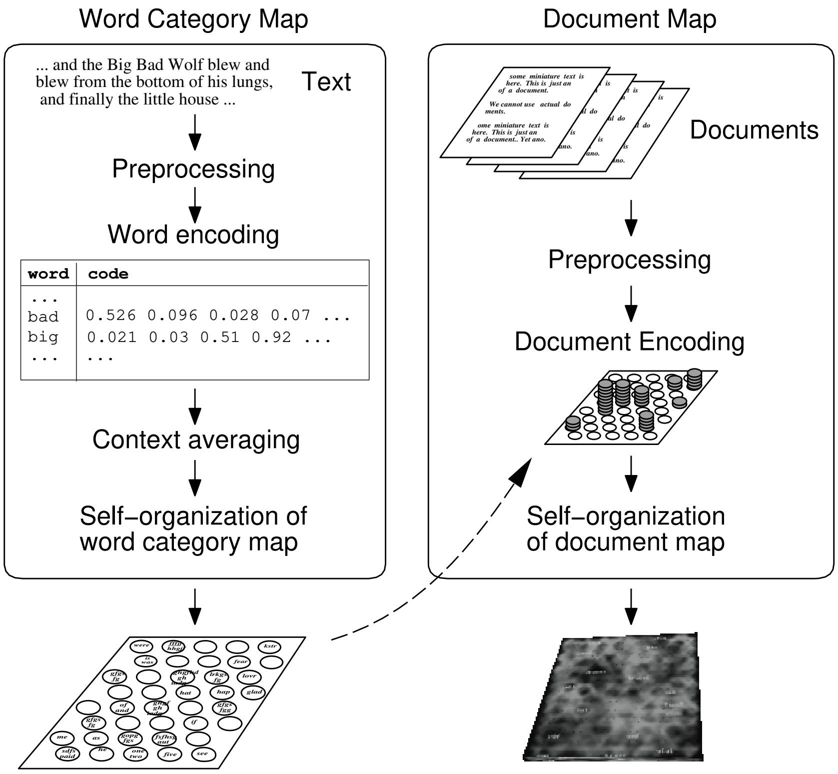

Figure 2: The basic architecture of the WEBSOM method. The document map is organized based on documents encoded with the word category map. Both maps are produced with the SOM algorithm. When the maps have been constructed, the processing of new documents is much faster.

formation of the document map.

2.1 Preprocessing

Before application of the Self-Organizing Map to a document collection some non-textual information (e.g., ASCII drawings and automatically included signatures) are removed. Numerical expressions and special codes are treated with heuristic rules.

To reduce the computational load the words that occur only a few times (say, less than 50 times) in the whole text corpus are neglected and treated as empty slots.

In order to emphasize the subject matters of the articles and to reduce erratic variations caused by the different discussion styles, common words that are not supposed to discriminate any discussion topics are discarded from the vocabulary.

2.2 Word category map

The word category map is a “self-organizing semantic map” [8] that describes relations of words based on their averaged short contexts. The ith word in the sequence of words is represented by an n-dimensional real vector x_i with random-number components. The averaged context vector of this word reads

$$X(i) = \begin{bmatrix} \mathrm{E}\{x_{i-1}|x_i\} \\ \varepsilon x_i \\ \mathrm{E}\{x_{i+1}|x_i\} \end{bmatrix} , \qquad (1)$$

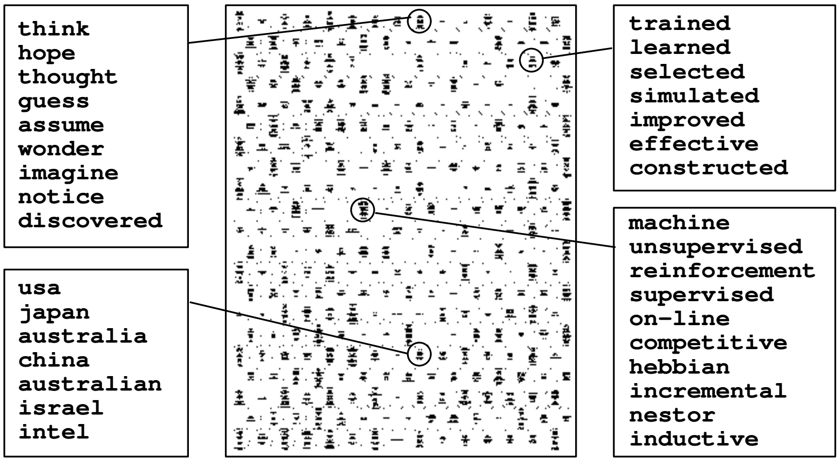

Figure 3: Examples of some clear "categories" of words on the word category map of the size of 15 by 21 nodes. The word labels of the map nodes have been shown with a tiny font on the map grid, and four nodes have been enlarged in the insets.

where E denotes the estimate of the expected value evaluated over the text corpus, and ε is a small scalar number. Now the $X(i) \in \Re^{3n}$ constitute the input vectors to the word category map. In our experiments $\varepsilon = 0.2$ and $n = 90$. The training set consists of all the $X(i)$ with different x_i.

The SOM is calibrated after the training process by inputting the $X(i)$ once again to the word category map and labeling the best-matching nodes according to symbols corresponding to the x_i parts of the $X(i)$. Usually interrelated words that have similar contexts appear close to each other on the map. Each node may become labeled by several symbols, often synonymous or belonging to the same closed category, thus forming "word categories" in the nodes. Sample categories are illustrated in Fig. 3.

2.3 Document map

The documents are encoded by mapping their text, word by word, onto the word category map, whereby a histogram of the "hits" on it is formed. To reduce the sensitivity of the histogram to small variations in the document content, the histograms are "blurred" using a Gaussian convolution kernel with, e.g., the full width at half maximum of two map spacings on the word category map consisting of 15 by 21 nodes. Such "blurring" is a commonplace method in pattern recognition for achieving invariance to small variation in the input, and is justified also here, because the map is ordered. The document map is then formed with the SOM algorithm using the histograms as "fingerprints" of the documents. To speed up the computation, the positions of the word labels on the word category map may be looked up by hash coding.

The document map has been found to reflect relations between newsgroup articles; similar articles tend to occur near each other on the map. Not all nodes are well focused on one subject only, however. While most discussions seem to be confined into rather small areas on the map, the discussions may also overlap. The visualized clustering tendency of the "digital library" is presented with the gray scale on the document map. Light areas correspond to clusters whereas transient areas between clusters are dark.

Computational speedups have been developed for creating very large maps [4]. In the largest experiments so far the word category map contained 315 units with 270 inputs each, and the document map had 104 040 neurons with 315 inputs each. The number of documents used for training the map in the experiment was 1 124 134.

3 Applications

The WEBSOM method is readily applicable to any kind of collection of textual documents. It is especially suitable for exploration tasks in which the users either do not know the domain very well, or they have only a limited idea of the contents of the full-text database being examined. With the WEBSOM, the documents are ordered meaningfully according to their contents. Maps also help the exploration by giving an overall view of what the information space looks like.

In addition to exploration, the WEBSOM may also be used for content-directed document search. Any new document may be mapped onto the document display. The position of the new document on the document map provides a starting point for exploring related documents in the nearby areas. The first version of this feature has recently been implemented. The result of a sample query is presented in Fig. 4.

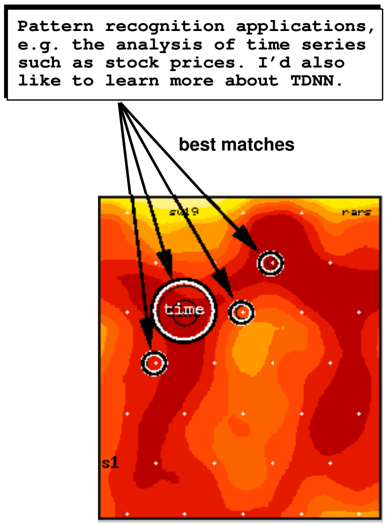

Figure 4: The result of a content-addressable search. The document has been positioned on a map that contains discussions on artificial neural networks. The area that was found is related to time-series prediction. The best-matching unit on the map is indicated by the largest circle.

Previously the SOM has been utilized, e.g., by Lin et al [6] to form a map based on titles of scientific documents. Scholtes has developed, based on the SOM, a neural filter and a neural interest map for information retrieval [9]. Merkl [7] has clustered textual descriptions of software library components. In comparison, one of the novel features of the WEBSOM method is the idea of applying the SOM algorithm twice: first for word category analysis and second for creating document maps, based on the first analysis.

In the World Wide Web, one application of the WEBSOM method could be ordering of home pages instead of the newsgroup articles. Also electronic mail messages could be automatically positioned on a suitable map according to personal interests. Relevant areas and single nodes on the map can be used as "mailboxes" into which specified information is automatically gathered (Fig. 5). The method could also be used to organize official letters, personal files, library collections, and corporate full-text databases. Administrative or legal documents may be difficult to locate by traditional information retrieval methods, because of the specialized terminologies used. For instance, the product developers of a company are likely to express themselves in different terms and paraphrases than the marketing staff. The category-based and redundantly encoded approach of the WEBSOM is expected to alleviate the terminology problem. We can also foresee the use of the method in the organization of conferences, in the automatic specification of sessions according to similar topics of the papers.

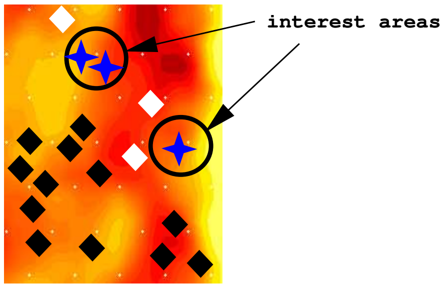

Figure 5: Schematic illustration of a WEBSOM document map used as a filtering tool. The circles denote the user's interest areas. The stars inside the circles are documents that would be selected by the system automatically. Those documents could, for instance, be instances of interesting electronic mail or articles from a news supplier. By virtue of the SOM, related documents — the white diamonds — can also be noticed and checked easily.

References

[1] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen. Newsgroup exploration with WEBSOM method and browsing interface. Technical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, January 1996. WEBSOM home page (1996) available at http://websom.hut.fi/websom/.

[2] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen. Creating an order in digital libraries with self-organizing maps. In *Proc of World Congress on Neural Networks (WCNN-96)*, 1996.

[3] T. Kohonen. *Self-Organizing Maps.* Springer, Berlin, 1995.

[4] T. Kohonen, S. Kaski, K. Lagus, and T. Honkela. Very large two-level som for the browsing of newsgroups. In *Proc. of International Conference on Artificial Neural Networks*, 1996.

[5] K. Lagus, T. Honkela, S. Kaski, and T. Kohonen. Self-organizing maps of document collections: A new approach to interactive exploration. In *Proc. of the 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD-96)*, Menlo Park, California, 1996. AAAI Press.

[6] X. Lin, D. Soergel, and G. Marchionini. A self-organizing semantic map for information retrieval. In *Proceedings of the 14th Annual International ACM/SIGIR Conference on Research & Development in Information Retrieval*, pages 262–269. 1991.

[7] D. Merkl and A. M. Tjoa. The representation of semantic similarity between documents by using maps: Application of an artificial neural network to organize software libraries. In *Proceedings of the General Assembly Conference and Congress of the International Federation for Information and Documentation, FID'94*, 1994.

[8] H. Ritter and T. Kohonen. Self-organizing semantic maps. *Biological Cybernetics*, 61:241–254, 1989.

[9] J. C. Scholtes. Unsupervised learning and the information retrieval problem. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN'91*, pages 18–21, Piscataway, NJ, 1991. IEEE Service Center.